HIGH SCHOOL EDITION
Journal of Student Research

Volume 14 Issue 1 (2025)

Predicting Enzyme Commission Numbers Using Recurrent Neural Networks with Amino Acid Sequence Shift and Consistency Loss

Danny Paik[1] and Giselle Gomes#

[1]Bergen County Academies, USA
#Advisor

ABSTRACT

Countless species of plants have been employed for the development of medications for diseases in all fields, ranging from viruses, cancer, diarrhea, and various skin diseases. However, the percentage of plant-based new molecular entity products compared to all types of medications has dropped by more than 50% since 1950, a major reason being the lack of information scientists and industries have for countless plant species. The majority of botanical medicines have been created off of plants that have already been known for their medicinal properties for centuries and have been utilized by specific cultures to treat illnesses. Therefore, the process of developing a plant-based medicine without much information about the plant can be incredibly expensive and time-consuming, leading to many industries avoiding doing so. Yet, plant-based drugs have been shown to have fewer adverse effects while maintaining high potency in comparison to synthetic drugs, making them a crucial part of medicine. Ergo, I proposed a method to accurately predict the Enzyme Commission (EC) number of plant genomes, allowing scientists to gain an understanding of the functionality of the genes in various plant species. The amino acid sequences will be represented on a 2D matrix, and I proposed a method to shift the matrix multiple times followed by stacking these respective matrices so the Convolutional Neural Network (CNN) can analyze a larger part of the amino acid sequence. Furthermore, EC numbers are a four-digit number sequence that categorizes all enzymes based on the type of chemical reactions they engage in. As the model will be predicting a sequence of numbers, I used a Recurrent Neural Network (RNN) to preserve the context in the sequence. Moreover, I proposed a method of adding a classifier for the prediction of the first digit of the EC number to decrease the sequential error.

Introduction

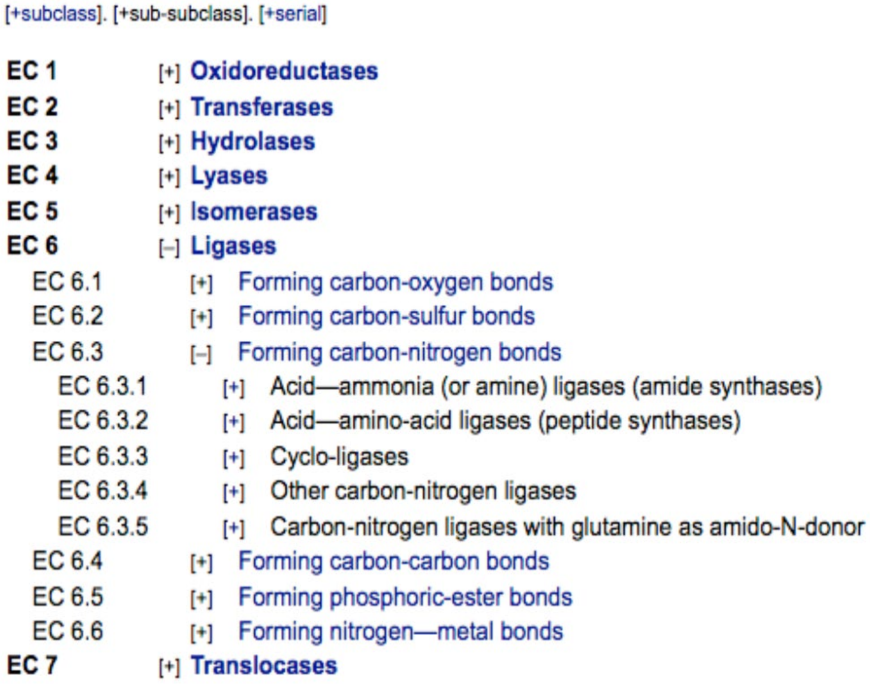
Natural medication, or medicine derived from natural products such as mammalians, bacteria, fungi, and plants, has been utilized by people for millennials. Notably, willow bark extract has been used as an anti-inflammatory and pain reliever for thousands of years (Shara and Stohs 2015). Such plant-based botanical drugs have played an invaluable role in the fields of drug discovery and medication, developing drugs that save the lives of countless people to this day such as Aspirin, Tamiflu, and Paclitaxel. Many botanical drugs have been developed through the existing knowledge of the medical capacities of certain plants that have been used for prolonged periods throughout history. For example, artemisinin was discovered by searching through ancient Chinese herb recipes and discovering the potent antimalarial substance in sweet wormwood.

This led to a revolutionary change in the combat against malaria, eventually being acknowledged with a Nobel Prize in medicine in 2015 (Su and Miller 2015). By already having qualitative data on these plants over decades if not centuries, many plant-based drugs have high chances of success and approval in earlier stages of clinical trials. However, the rate of success in finding a plant and synthesizing a drug is extremely low and also requires considerable amounts of time and resources in testing (Patridge et al. 2016). Thus, the percentage of plant-based New Molecular

Entity (NME) products has dropped by more than 50% since 1950, now barely lingering under 8.7% of all NME products (Patridge et al. 2016). Requiring extensive prior knowledge of a plant in order to develop a drug further has been a major challenge, stunting the development of botanical drugs over the years. There are almost 400,000 plant species in the world, with only 31,000 having a documented use, roughly having a usage rate below 8%. Plant-based drugs are likely to have fewer adverse effects and are much more environmentally sustainable than synthetic drugs, making them an incredible asset in the field of medicine.

To expedite the process of plant-based drug discovery and development, it is important to further analyze the properties of potential plants for drug discovery, as it is clear that there is a lack of knowledge and usage of plants that are still open for examination. To do so, I will utilize a machine learining based high-throughput screening of the molecular structures of the phytochemicals of plants and further analyze the enzyme commission number of the proteins in the plants to offer a deeper understanding of potential botanical medicine.

Enzyme Commission Number



[+subclass]. [+sub-subclass]. [+serial]

EC 1	[+]	Oxidoreductases
EC 2	[+]	Transferases
EC 3	[+]	Hydrolases
EC 4	[+]	Lyases
EC 5	[+]	Isomerases
EC 6	[–]	Ligases
EC 6.1	[+]	Forming carbon-oxygen bonds
EC 6.2	[+]	Forming carbon-sulfur bonds
EC 6.3	[–]	Forming carbon-nitrogen bonds
EC 6.3.1	[+]	Acid—ammonia (or amine) ligases (amide synthases)
EC 6.3.2	[+]	Acid—amino-acid ligases (peptide synthases)
EC 6.3.3	[+]	Cyclo-ligases
EC 6.3.4	[+]	Other carbon-nitrogen ligases
EC 6.3.5	[+]	Carbon-nitrogen ligases with glutamine as amido-N-donor
EC 6.4	[+]	Forming carbon-carbon bonds
EC 6.5	[+]	Forming phosphoric-ester bonds
EC 6.6	[+]	Forming nitrogen—metal bonds
EC 7	[+]	Translocases

Figure 1. NC-IUBMB Enzyme List Showing Hierarchical Classification By EC Numbers (McDonald and Tipton 2023)

Enzymes are essential proteins for living organisms that act as catalysts to accelerate biochemical reactions in the body (Robinson 2015). With over 4,000 enzymes responsible for chemical reactions in living organisms, it is necessary to organize the properties and functionalities of these enzymes, Enzyme Commission (EC) numbers have been assigned to enzymes to mark their functions, helping the process of inferring and analyzing biological and genetic properties of new species through the study of their protein sequences. Along with the findings of countless enzymes, there are currently over 565,928 manually curated protein sequences and over 225,013,025 that are computationally annotated by EC numbers. EC numbers are composed of four digits representing each of the following: class, subclass, sub-subclass, and serial number. (Han et al. 2023).

Each of these digits is used to describe which reactions the enzymes act as catalysts. The first digit which shows the class, identifies the enzyme as one of the following 7 groups, with Translocases being a relatively new class: Oxidoreductases (EC 1), Transferases (EC 2), Hydrolases (EC 3), Lyases (EC 4), Isomerases (EC 5) Ligases (EC 6) and Translocases (EC 7) (McDonald and Tipton 2023). The sub-class and sub-subclass further help classify the enzyme based on reactions, and the serial number is a unique number given to enzymes to identify them. This hierarchical classification system of enzymes allows for further analysis and inference of functions of genes and properties of genomes by looking at the protein sequences (Kim et al. 2023).

Proposed Enzyme Commission Number Prediction Network

The proposed method comprises three modules: preprocessing, feature extraction, and enzyme commission (EC) number prediction. The preprocessing module converts the amino acid sequence into a 2-D matrix embedding representation, preparing it for input into the convolutional neural network (CNN)-based feature extraction module. The CNN-based feature extraction module takes the 2-D matrix embedding as input and generates feature maps that mathematically represent the features of the input amino acid sequence. These extracted feature maps are then fed into the EC number prediction network.

Preprocessing

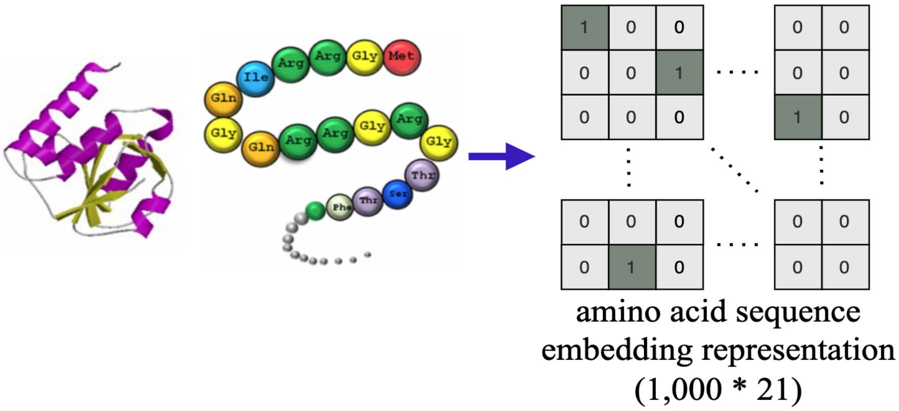


Figure 2. Preprocessing (embedding representation of amino acid seqeunce)

Enzymes are composed of an extended chain of countless amino acids. In order to thoroughly analyze the amino acid sequence that represents the enzyme, it is crucial to embed the amino acids in a matrix that can be further studied by a neural network. In order to do so, the matrix will be composed of 1,000 rows for each individual amino acid in the sequence, and 21 columns with each column representing one of the 21 types of amino acids. However, the Convolutional Neural Network (CNN) observing this matrix will only be able to view a limited number of rows at a time due to its shape, translating to only a handful of amino acids being viewed simultaneously during analysis. Therefore, to resolve this issue, I propose a method of “shifting” which shifts the rows of the matrix down by one for each execution. This will be done by finding an optimized number of ***N***, to create ***N*** differently shifted matrices through ***N*** times of processing.

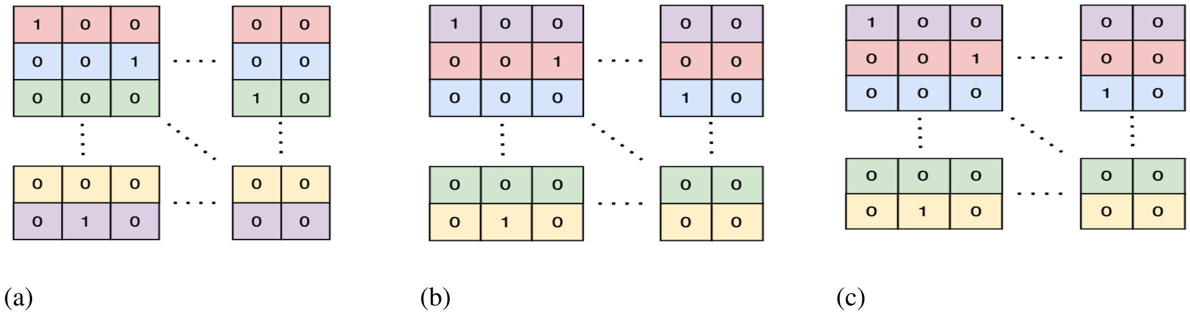


Figure 3. Proposed amino acid sequence shifting
(a): original embedding representation, (b): one offset embedding representation, and (c): two offset embedding representation

Feature Extraction

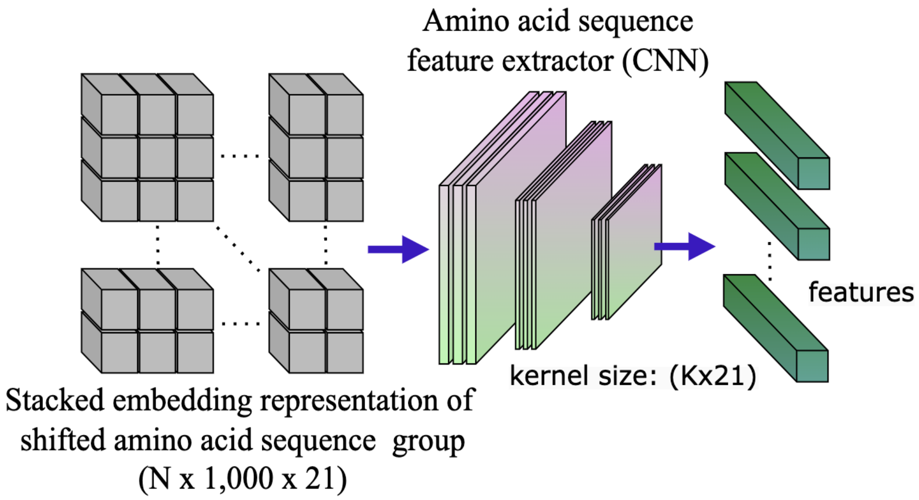

Amino acid sequence feature extractor (CNN)
features
kernel size: (Kx21)
Stacked embedding representation of shifted amino acid sequence group (N x 1,000 x 21)

Figure 4. Feature extraction of stacked embedding representation

The N different matrices would then be stacked together, transforming the originally 2-dimensional $1{,}000 \times 21$ matrix into a 3-dimensional $N \times 1{,}000 \times 21$ matrix. This allows the CNN to consume a greater multitude of information consisting of considerable amino acids, thus leading to improved accuracy and context preservation of the amino acids. The stacked matrix representing the shifted amino acid sequence group will then go through an amino acid sequence feature extractor which creates the feature maps of the amino acid sequence in the appropriate length to be analyzed thoroughly.

Enzyme Commission Number Prediction

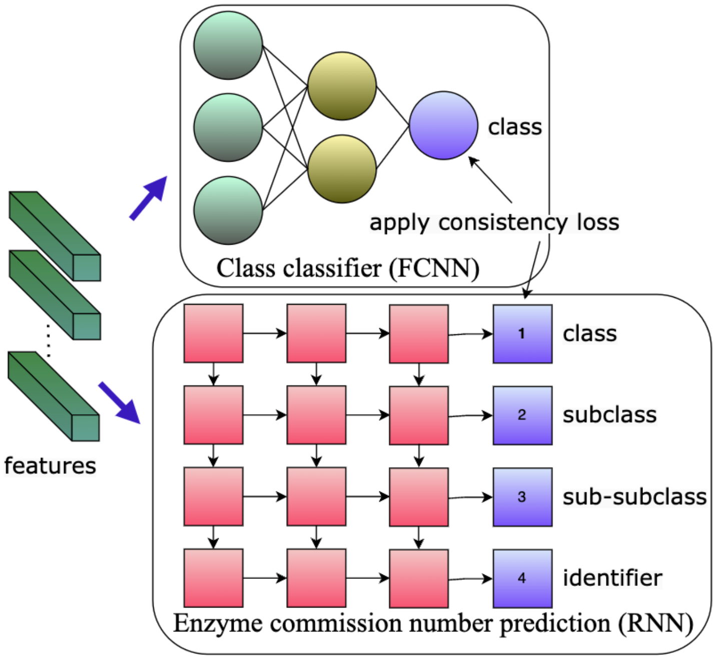


Figure 5. LSTM-based enzyme commission number prediction

For the architecture of the model to predict the EC number of an enzyme, I will be using a Recurrent Neural Network (RNN) which is a neural network that goes through multiple steps of preserving sequential data. It accounts for all the values computed before, allowing it to robustly preserve the context in a sequence. EC numbers are a sequence of 4 numbers identifying the class, subclass, sub-subclass, and identifier, making it optimal to use an RNN for prediction. On the other hand, one of the major flaws of an RNN is the fact that it uses previous results for the computation of the sequentially following values.

Thus, an error in the earlier stage of an RNN is amplified throughout the sequential computation process, making the accuracy of the first calculation invaluable for the overall accuracy of the model. To increase the accuracy of the whole model, I implemented a Fully Connected Neural Network (FCNN) to function as a class classifier of the enzyme. An enzyme only has 7 different classes: Oxidoreductases (EC 1), Transferases (EC 2), Hydrolases (EC 3), Lyases (EC 4), Isomerases (EC 5) Ligases (EC 6), and Translocases (EC 7). Therefore, it would be reasonable to add a class classifier to increase the accuracy of predicting the class, and therefore predicting the rest of the sequence much more accurately as well.

Equation 1. Cross entropy loss function for Class Classifier (first digit of EC number)

$$L_{FCNN} = -\sum_i gt_i^1 \times log_e(\hat{p}_i)$$

Equation 2. Cross entropy loss function for LSTM-based Classifier

$$L_{RNN} = -\frac{1}{N}\sum_n \sum_i gt_i^n \times log_e(\hat{o}_i^n)$$

Equation 3. Consistency loss function

$$L_{consistency} = \sum_i gt_i^1 \times (\hat{p}_i - \hat{o}_i^1)$$

The loss of the FCNN will be measured via a cross-entropy loss function to determine the accurate class of the enzyme. This will be done for every row of the matrix and the sum will account for the total loss of the class classifier. The ground truths used for all of these functions will be 1 only for the correct corresponding EC number. A cross-entropy loss function will also be utilized to compute the total loss of the RNN down by the rows of the original matrix and then the width N of the stack. The consistency is computed by taking the difference between the probability from the classifier and the first row values of the RNN. This will evaluate the consistency of the classification of the class and the sequentially following predicted values.

Experimental Results

Dataset and Inference Metrics

The dataset consists of a training set to sort E.C. numbers in genetic sequences with a size of 7,690,892, of which 5,760,151 (74.90 %) are non-enzymatic and 1,930,741 (25.10 %) are enzymatic. This dataset is based on plant genomes, therefore making it crucial in predicting the functions of plants by analyzing their E.C. numbers. My model had a peak accuracy of 0.7151, when having 4 LSTM layers and respectively had its highest F1-score with the same structure, achieving an F1-score of 0.7129. Moreover, in order to assess the effectiveness of the shifting and FCNN, I tested two more models, one with neither a shifting in stacking nor an FCNN, and one with shifting but no FCNN.

Evaluation Protocol

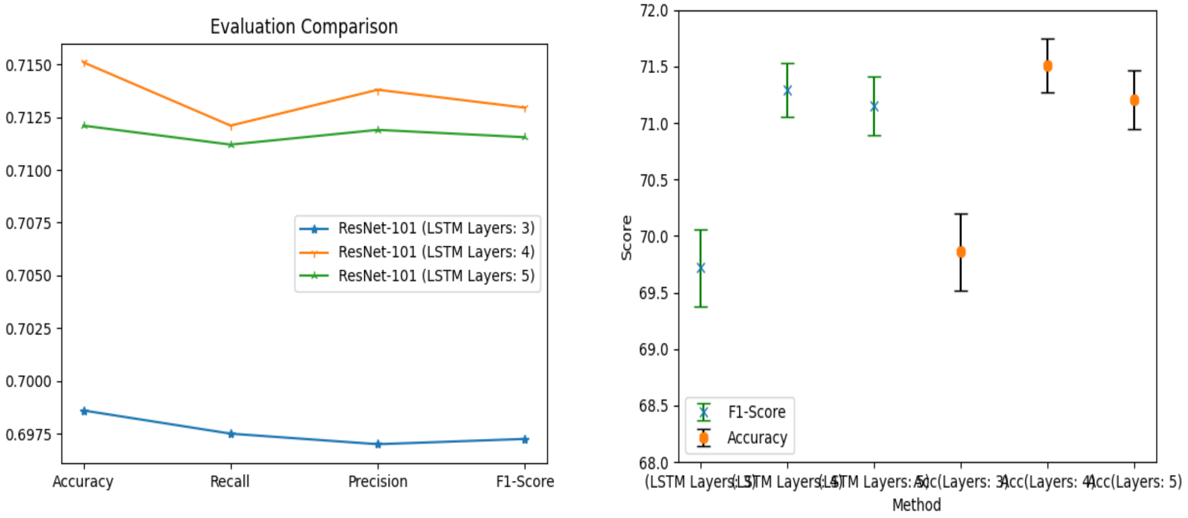
The dataset was split such that 80% of the dataset was used to train the model and the remaining 20% was used for testing with a 5-fold cross-validation. This means that the dataset was split into 5 sets of samples and the training was done 5 times, rotating by which set will be used for testing. For each rotation, a varying number of K feature extractors were selected and the F1-score, which is the harmonic mean of the recall and precision, was calculated for each of the trained structures. For the feature extractors, I experimented with a different number of LSTM layers for ResNet-101. On every rotation, 4 sets were used for training and the remaining 1 set was used to validate the trained model. The mean F1-score was then calculated along with the standard deviation and plotted along with the accuracy of the model.

Evaluation Results

Table 1. Evaluation results with different layer setup

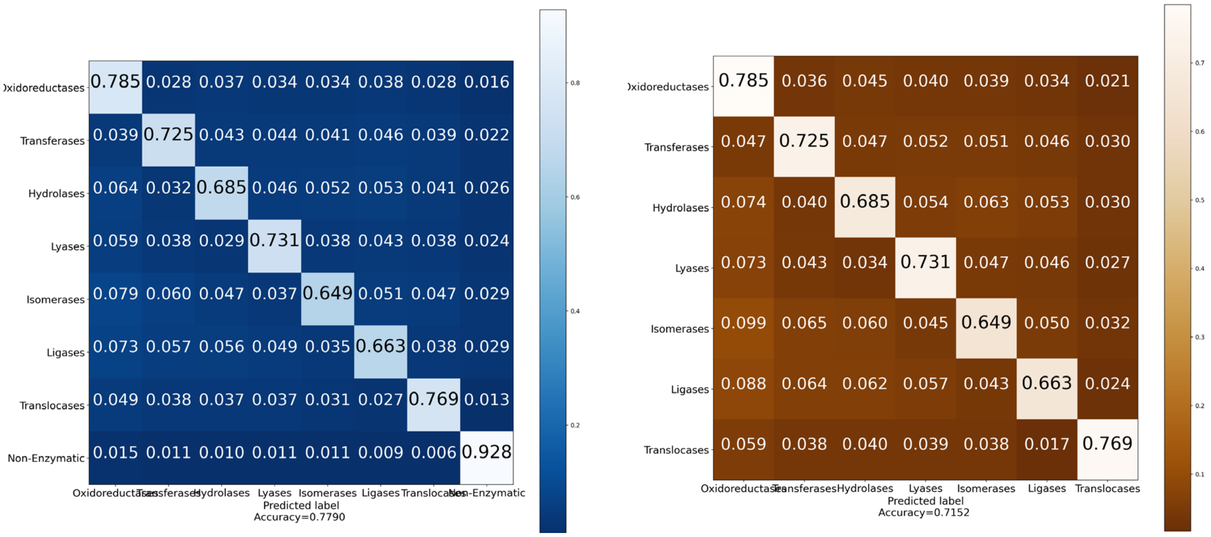
	Accuracy	Recall	Precision	F1-Score
ResNet-101 (LSTM Layers: 3)	0.6986	0.6975	0.6970	0.6972
ResNet-101 (LSTM Layers: 4)	0.7151	0.7121	0.7138	0.7129
ResNet-101 (LSTM Layers: 5)	0.7121	0.7112	0.7119	0.7115

First, in the case of the feature extractors, the accuracy, and F1-scores, when there were 3, 4, and 5 LSTM layers, were (0.6986, 0.6972), (0.7151, 0.7129), and (0.7121, 0.7115) respectively. It showed to have the highest accuracy, recall, and precision when there were 4 layers, whereas the accuracy and F1-score started to decline when there were 5 layers due to overfitting.



(a) (b)

Figure 6. (a): Evaluation results with different layer setup and (b): K-fold cross validation result



(a) (b)

Figure 7. (a): Confusion matrix of EC number (7 classes) and Non-Enzymatic classification and (b): (a): Confusion matrix of EC number (7 classes)

Moreover, for the ablation models where one had no shifting nor an FCNN, and one where there was shifting but no FCNN, their F1-scores were 0.6248 and 0.6721 respectively, highlighting the significance of the shifting and

FCNN for the model in terms of precision and recall. Therefore, the proposed model of shifting the embedded amino acid sequence representation when stacking, and the addition of a classifier for the class of the E.C. number proved to be crucial to accurately predict the E.C. numbers in the genome of plants.

Table 2. Ablation study result

Method	F1-Score
Ablation Model 1 [stacked(duplication) embedding + LSTM]	0.6248
Ablation Model 2 [stacked(shifted) embedding]	0.6721
Proposed [stacked(shifted) embedding + consistency loss]	0.7129

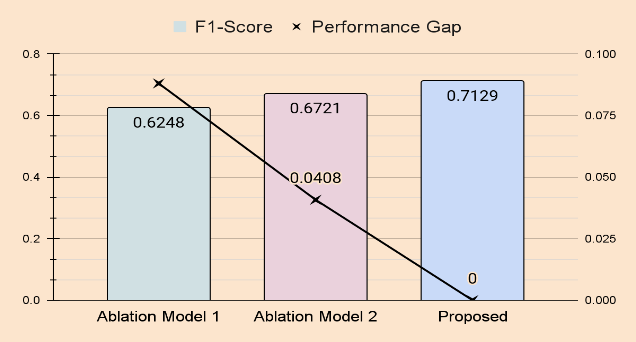


Figure 8. Ablation study result

Conclusion

This paper offers a novel method to predict the EC numbers in a plant genome, significantly boosting the efficiency for the development of potential plant-based medicine. There are over 400,000 plant species in the world and to develop botanical drugs or medications, typically, a long history with the use of a plant is needed to provide background knowledge to develop the medicine. EC numbers are numbers assigned to enzymes in order to categorize them based on their chemical reactions. Therefore, EC numbers in plant genomes offer a deep analysis of the characteristics and functions of the genes, making it vital for obtaining knowledge of shallowly studied plant species. I proposed a method to shift the embedded representation of the amino acid sequence when stacked in order for the CNN to analyze a longer portion of the amino acid sequence than before. Moreover, I added an FCNN to help improve the accuracy of the prediction for the class of the EC number because the model utilized an RNN which is very dependent on the first result which in this case is the first digit of the EC number; the class. This model proved to be highly viable with a peak accuracy of 0.7151 and an F1 Score of 0.7129. In the future, I intend to further improve the accuracy of the model and test the model on a larger variety of plant species to offer a multitude of information in the field of drug discovery, potentially aiding the development of more botanical drugs.

Acknowledgments

I would like to thank my advisor for the valuable insight provided to me on this topic.

References

AI Hub. (2024, Sep 11). “Plant functionality prediction genomic data”: AI Hub. https://www.aihub.or.kr/aihubdata/data/view.do?currMenu=115&topMenu=100&aihubDataSe=data&dataSetSn=71316

Arora, C., Verma, D. K., Aslam, J., & Mahish, P. K. (Eds.). (2023). Phytochemicals in Medicinal Plants: Biodiversity, Bioactivity and Drug Discovery. Walter de Gruyter GmbH & Co KG.

Han, S. R., Park, M., Kosaraju, S., Lee, J., Lee, H., Lee, J. H., ... & Kang, M. (2024). Evidential deep learning for trustworthy prediction of enzyme commission number. Briefings in Bioinformatics, 25(1), bbad401.

Kim, G. B., Kim, J. Y., Lee, J. A., Norsigian, C. J., Palsson, B. O., & Lee, S. Y. (2023). Functional annotation of enzyme-encoding genes using deep learning with transformer layers. Nature Communications, 14(1), 7370.

Kumar, A., P, N., Kumar, M., Jose, A., Tomer, V., Oz, E., ... & Oz, F. (2023). Major phytochemicals: recent advances in health benefits and extraction method. Molecules, 28(2), 887.

Mani, J. S., Johnson, J. B., Steel, J. C., Broszczak, D. A., Neilsen, P. M., Walsh, K. B., & Naiker, M. (2020). Natural product-derived phytochemicals as potential agents against coronaviruses: A review. Virus research, 284, 197989.

McDonald, A. G., & Tipton, K. F. (2023). Enzyme nomenclature and classification: The state of the art. The FEBS journal, 290(9), 2214-2231.

Robinson, P. K. (2015). Enzymes: principles and biotechnological applications. Essays in biochemistry, 59, 1.

Shara, M., & Stohs, S. J. (2015). Efficacy and safety of white willow bark (Salix alba) extracts. Phytotherapy Research, 29(8), 1112-1116.

Su, X. Z., & Miller, L. H. (2015). The discovery of artemisinin and the Nobel Prize in Physiology or Medicine.